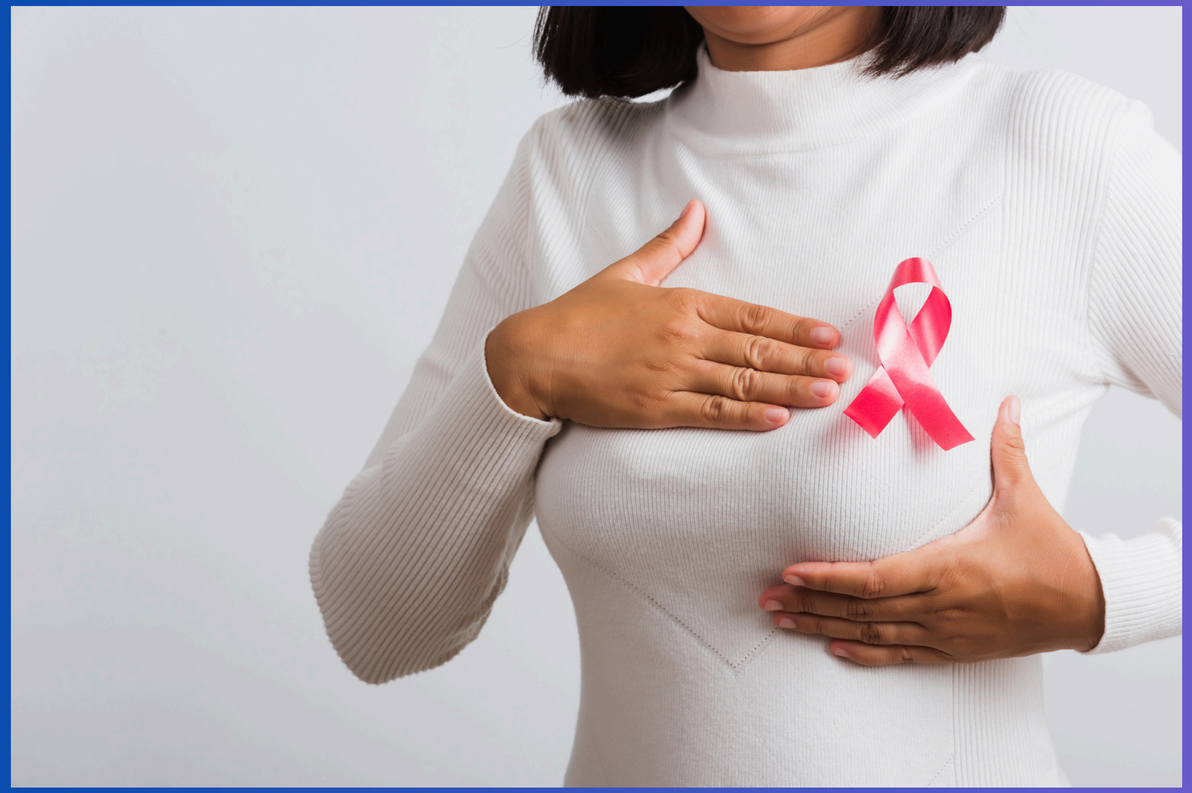


"Early detection saves lives, let's fight together."

DR. SARAH JOHNSON, BREAST CANCER RESEARCHER

HISTOPATHOLOGY CLASSIFICATION

Identification of breast cancer from histopathological images using Image Classification

Final Capstone Project

Team Members:

- **Andrew Mutuku**
- **Amina Saidi**
- **Wambui Githinji**
- **Winnie Osolo**
- **Joseph Karumba**
- **Margaret Njenga**

INTRODUCTION

- Breast cancer, the most common cancer in Kenya, is often diagnosed at advanced stages due to limited access to timely diagnosis and treatment.

- This project aims to develop a machine learning model for accurately classifying histopathological images of breast tumors, improving diagnostic accuracy, early detection, and treatment outcomes, especially in low- and middle-income countries like Kenya.

METHODOLOGY

INTRODUCTION

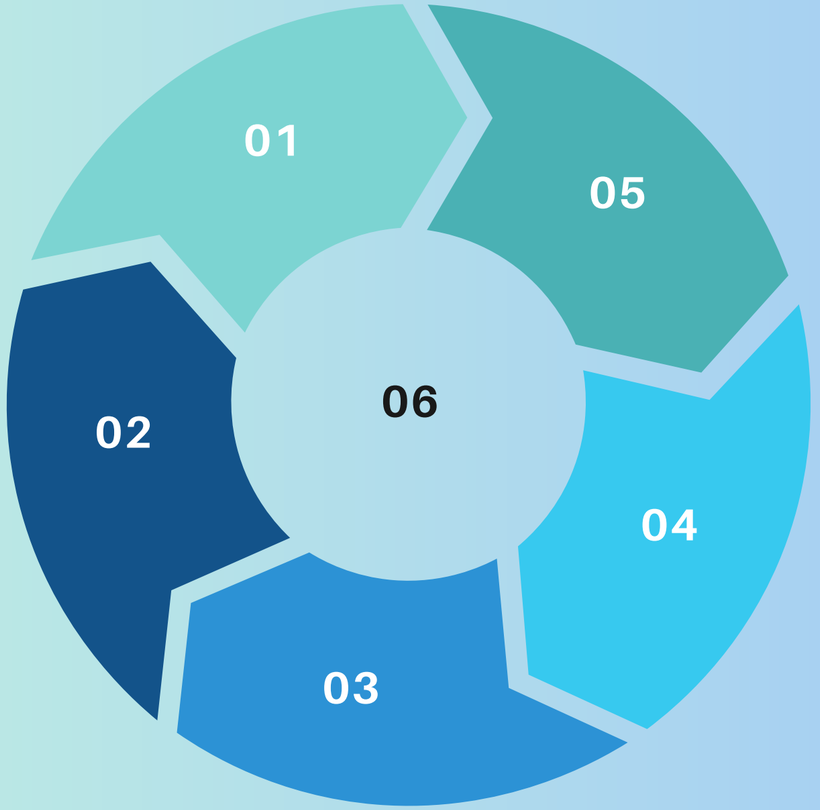
Business Problem and Project Objectives

DATA UNDERSTANDING AND DATA PREPARATION

Dataset, Image specifications and Classes

EXPLORATORY DATA ANALYSIS

Class distribution, sample images and observations



DATA PREPROCESSING & MODELING

Image resizing and normalization, Data Augmentation techniques, model choice, modification, training parameters, training process

EVALUATION & METRICS

Metrics, Detailed Evaluation, Perfomance Analysis, Key Findings, Implications

DEPLOYMENT

OBJECTIVES

1. Develop a robust image classification model to distinguish between benign and malignant breast tumor histopathological images
2. Enhance Diagnostic Accuracy through automated analysis.

RESEARCH QUESTIONS

How accurately can machine learning models classify benign and malignant breast tumor histopathological images?

How can machine learning improve the diagnostic process for breast cancer in low- and middle-income countries?

Which machine learning algorithms yield the highest precision and recall in classifying breast cancer images?

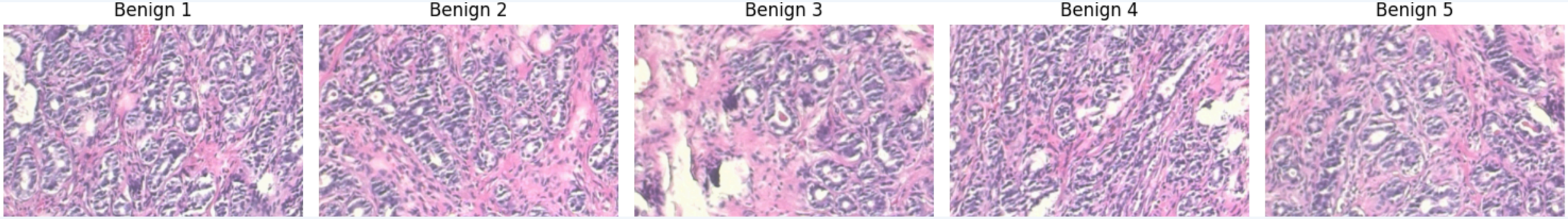
DATA UNDERSTANDING

- We sourced our data from the Breast Cancer Histopathological Image Classification (BreakHis) database, which contains 9,104 microscopic images of breast tumor tissue, including 2,480 benign and 5,429 malignant samples.

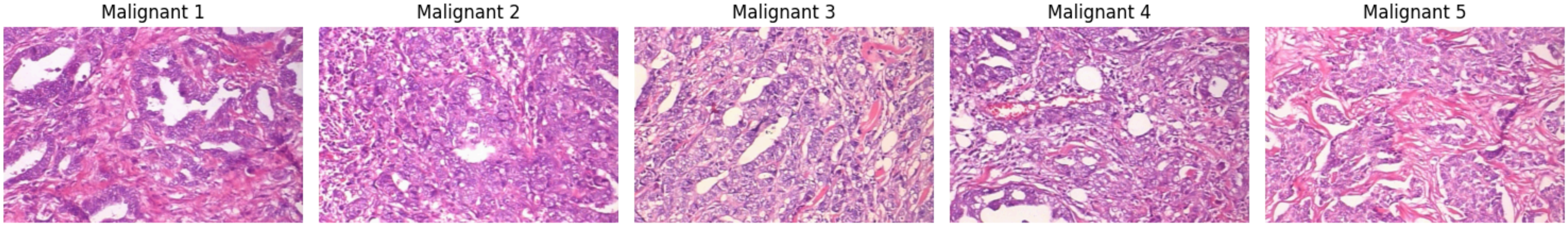
- In the Kenyan Context, most machines used in cancer detection are light microscopes which can focus images at either 40X magnification or 100X magnification.

- We opted to focus on these magnifications in our model. Each image file name stores essential information such as the method of procedure biopsy, tumor class (benign or malignant), tumor type, patient identification, and magnification factor.

BENIGN IMAGES



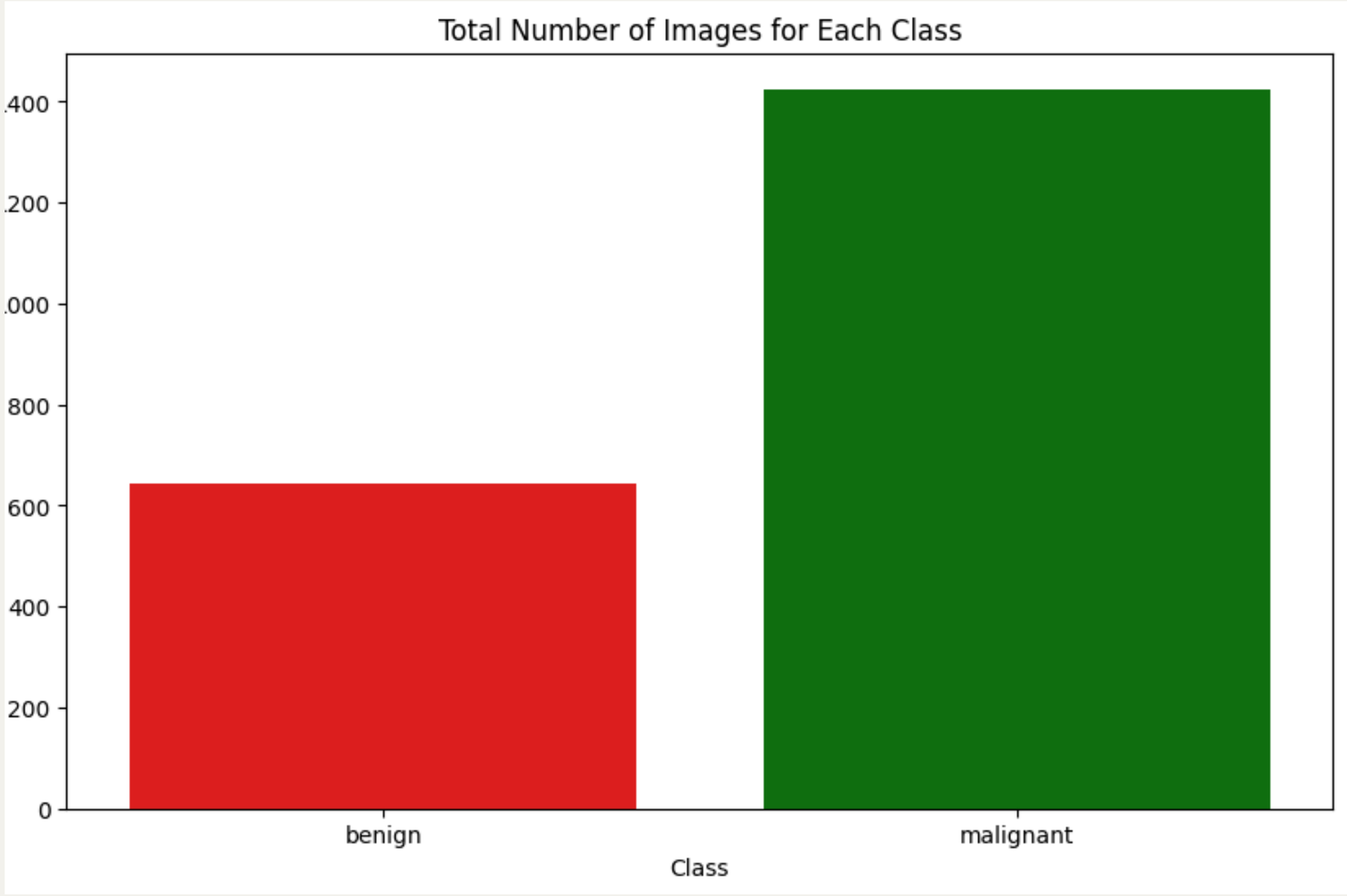
MALIGNANT IMAGES



The benign images show organized and well-differentiated tissue structures with uniform cell distribution, indicating non-cancerous growths. In contrast, the malignant images display disorganized cellular structures, increased density, and irregular boundaries, characteristic of invasive cancerous tissue. These visual differences help in distinguishing between benign and malignant breast tumors for accurate diagnosis and treatment planning.

CLASS DISTRIBUTION

The bar chart displays the total number of images categorized as either benign or malignant.



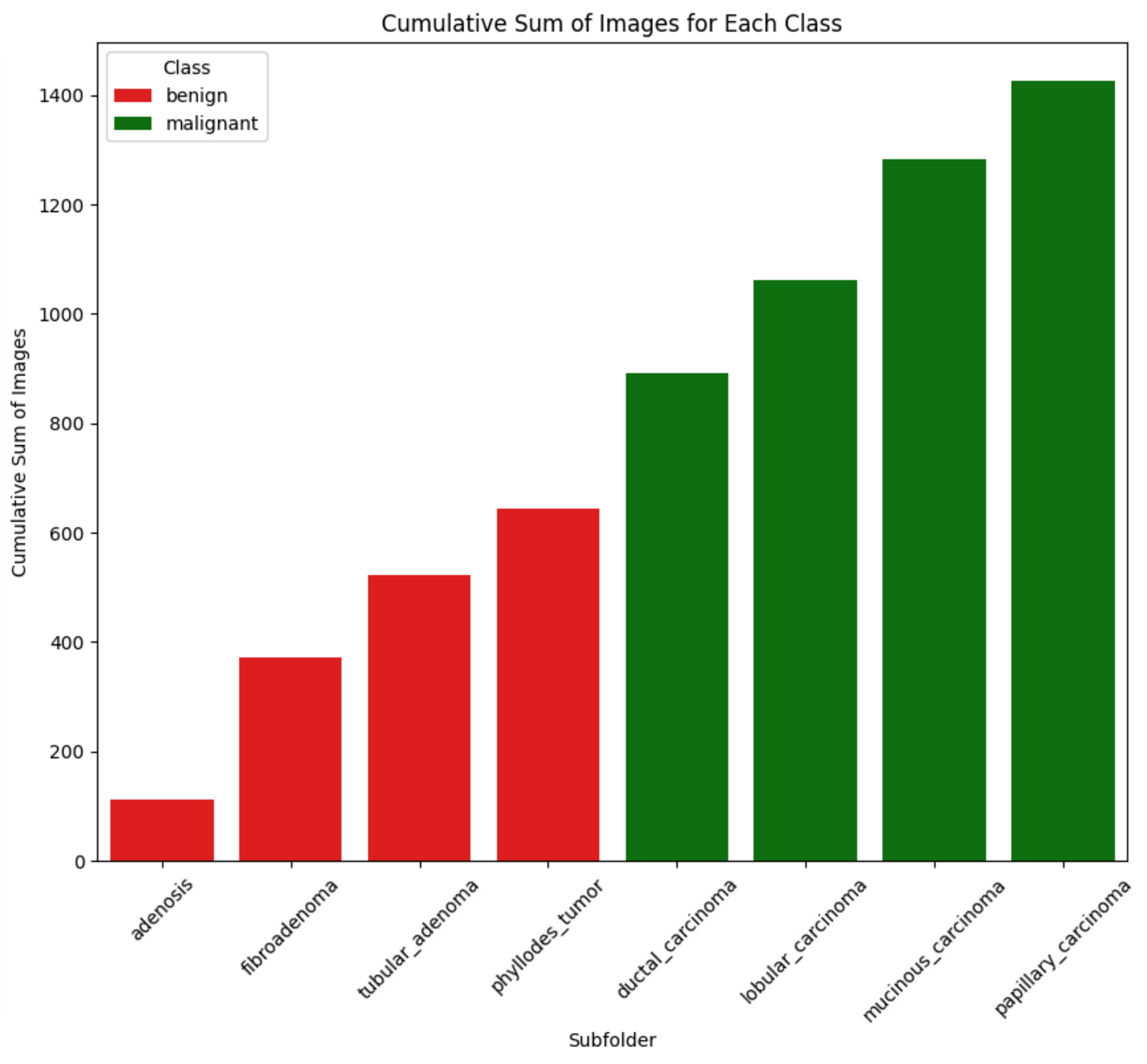
- It shows that there are significantly more malignant images (over 1400) compared to benign images (around 600).

- This indicates a class imbalance in the dataset, with malignant images being more than double the number of benign images.

CUMULATIVE SUM OF IMAGES

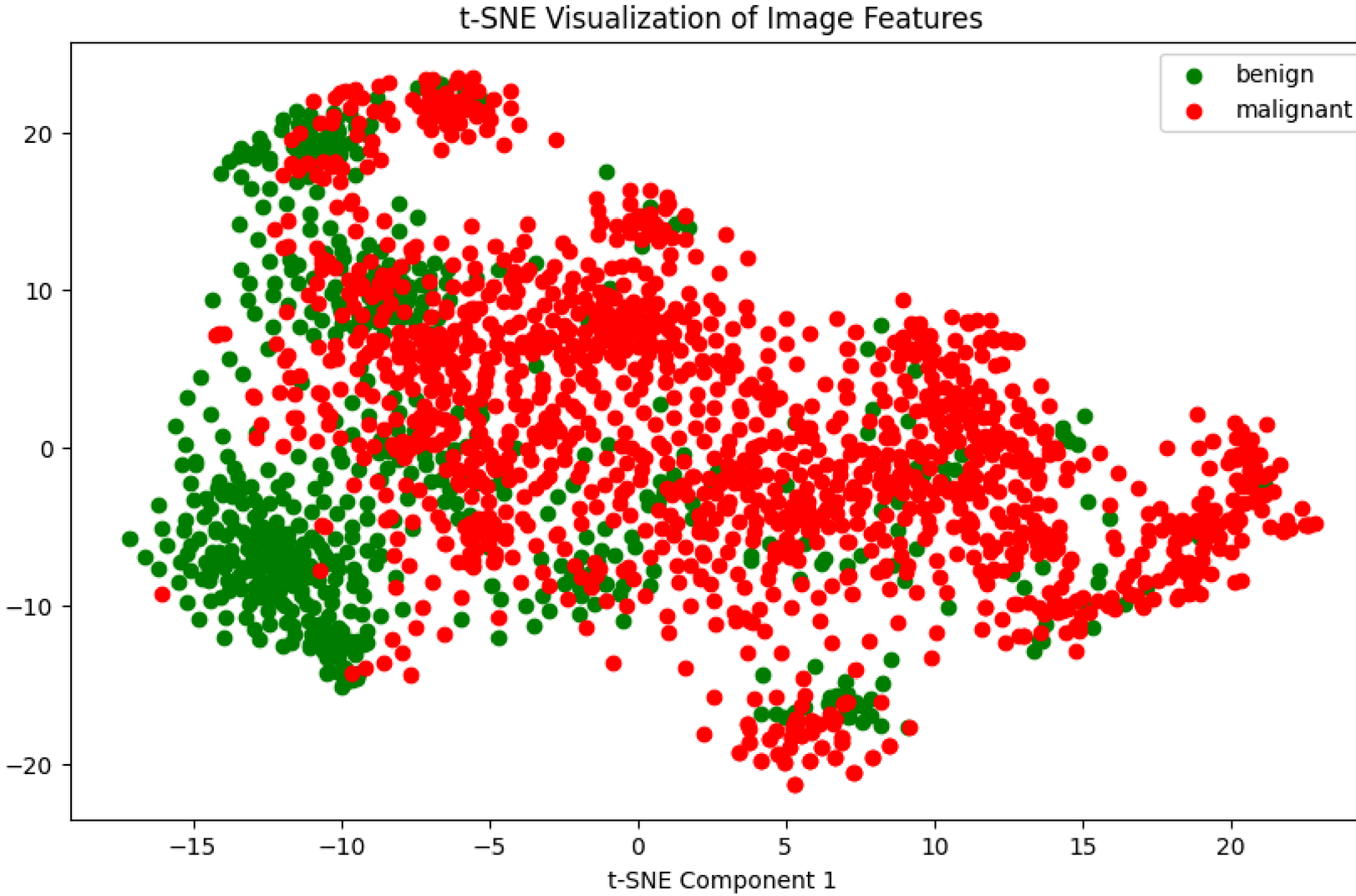
The bar chart shows the cumulative sum of images for benign and malignant classes across different subfolders.

- The highest number of images is in the "ductal_carcinoma" subfolder under the malignant class, with over 1400 images.
- The benign subfolders "phyllodes_tumor" and "adenosis" also have high cumulative sums, with around 600 images each.
- This indicates a significant variation in the number of images per subfolder, especially within the malignant class.



T-SNE VISUALIZATION

- The t-SNE plot visualizes the high-dimensional image features in a 2D space, differentiating between benign (green) and malignant (red) categories.
- The malignant images are spread more broadly across the plot, while the benign images form more distinct clusters, particularly on the left side.



MODELLING

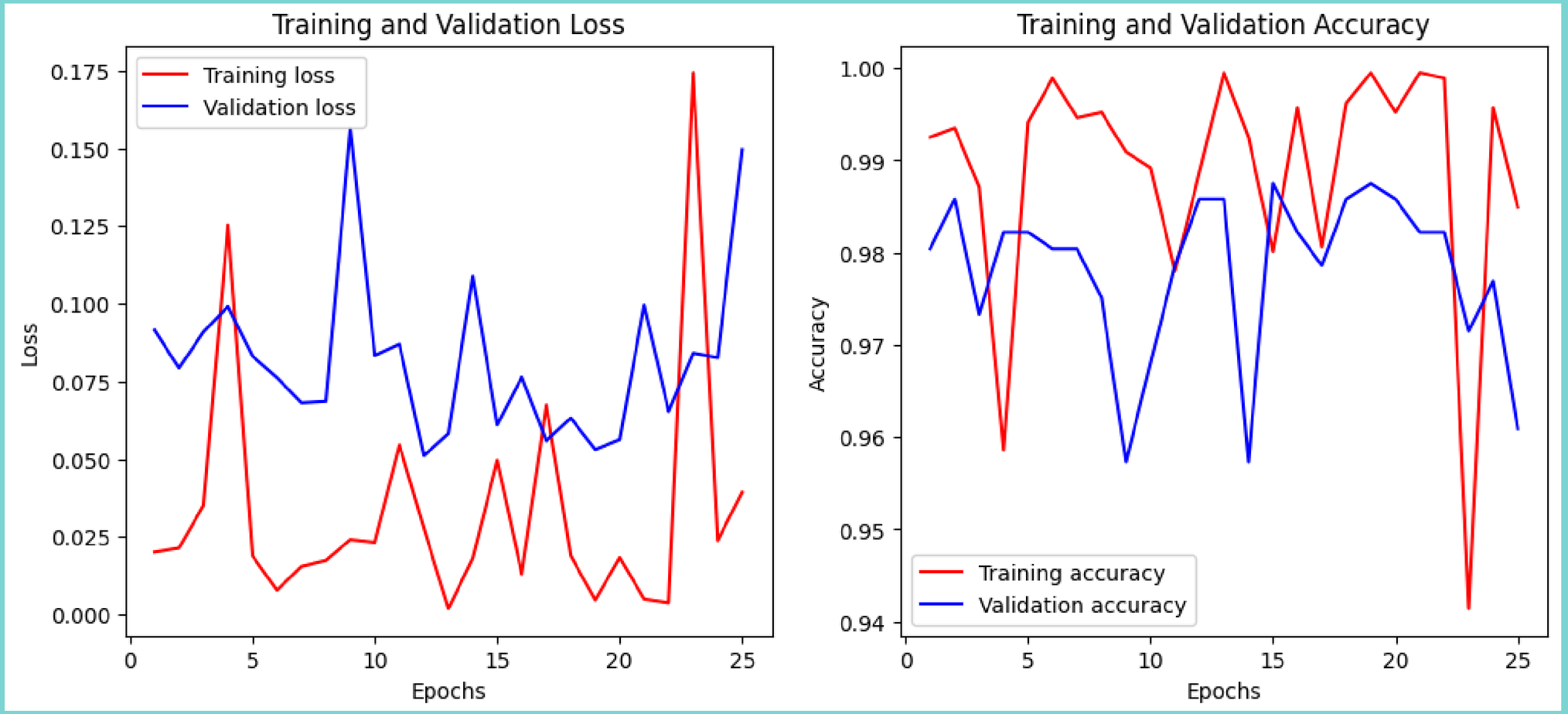
Overview

- We used a combination of DenseNet, CNN, and VGG16 models for our breast cancer image classification.
- Aim: Achieve high precision, recall , accuracy and robust performance in image classification.

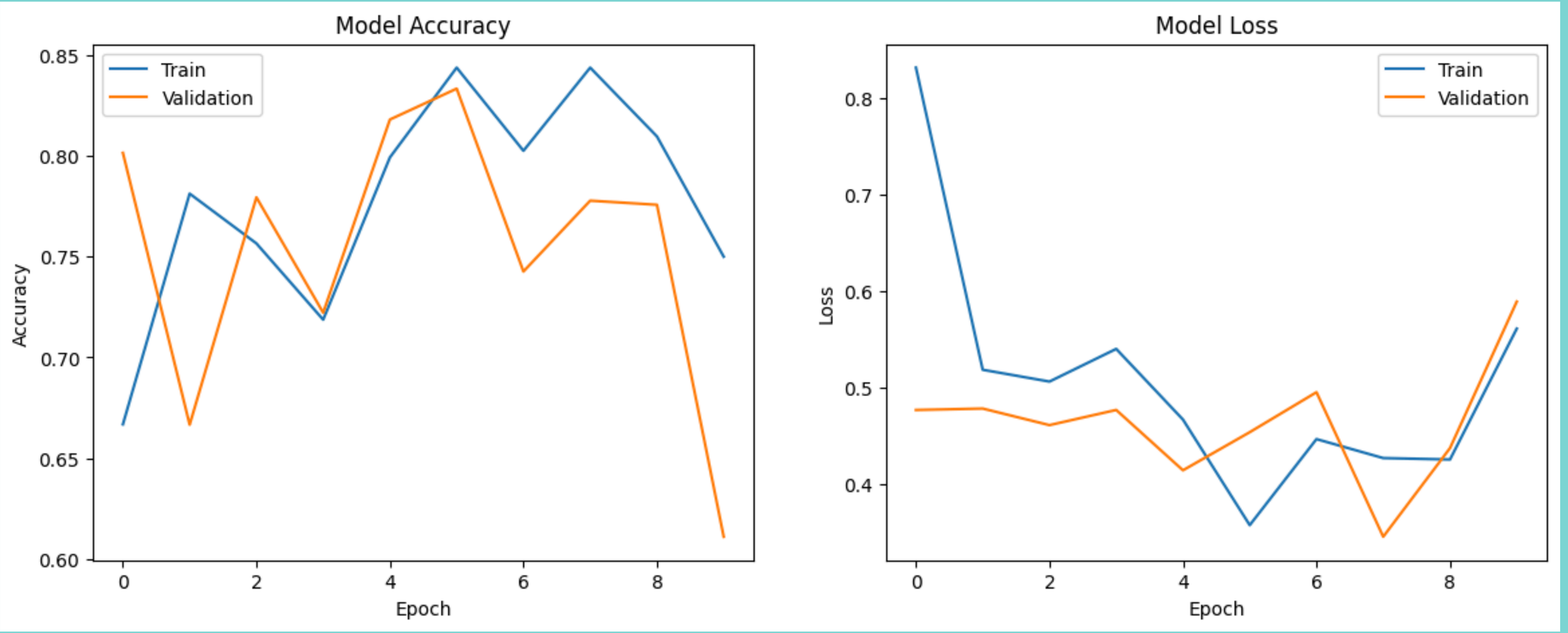
	Dense Net	CNN
Accuracy	0.9858	0.9048
Precision	0.985	0.904
Recall	0.9821	0.9048
F1 Score	0.9836	0.9037

Model Performance

Overall, DenseNet performs better across all metrics compared to the CNN but is indicative of overfitting, making CNN a better model for this task.



In the DenseNet, training loss is lower and the training accuracy is higher than the validation loss and accuracy, respectively, with noticeable high variance. This suggests that the DenseNet model may be overfitting, meaning it performs well on the training data but less reliably on new, unseen data.



The CNN model exhibits a more stable performance with higher validation accuracy compared to training accuracy and lower validation loss compared to training loss. This stability indicates that the CNN model generalizes better to new data,

RECOMMENDATIONS

TRAINING AND EDUCATION

IINTEGRATION INTO HEALTHCARE SYSTEMS

CREATING PUBLIC AWARENESS

CONCLUSION

- This project highlights how machine learning can improve breast cancer diagnosis in Kenya. By accurately classifying histopathological images as benign or malignant, the model reduces diagnostic errors and speeds up the process.

- This is crucial for addressing high breast cancer mortality rates in Kenya, where late-stage detection is prevalent.

- With a detection time of 45 seconds, the model aids in faster diagnoses, potentially shortening turnaround times and enhancing patient outcomes.

DEPLOYMENT

- The model was deployed using Streamlit, allowing medical practitioners to upload histopathological images for real-time diagnosis.

- Performance Monitoring: Implement logging and monitoring to track the model's performance and schedule periodic retraining with updated data.

NEXT STEPS

- Data expansion: Continuously expand and update the dataset with new histopathological images to improve the model's robustness and accuracy over time.
- Model Refinement: Explore advanced techniques such as transfer learning, fine-tuning, and additional ensemble methods to further enhance model performance.
- Integration with Clinical Systems: Develop interfaces for integrating the model with existing clinical systems to streamline workflow and facilitate real-time diagnosis.
- Continuous Monitoring and Feedback Loop: Establish a feedback loop with medical professionals to regularly refine the model. Utilize their insights to enhance predictions and update the model with new data over time.

THANK

YOU